Dam failure peak outflow prediction through GEP-SVM meta models and uncertainty analysis

Mohammad Nobarinia, Farhoud Kalateh, Vahid Nourani and Alireza Babaeian Amini

ABSTRACT

Accurate prediction of a breached dam’s peak outflow is a significant factor for flood risk analysis. In this study, the capability of Support Vector Machine and Kernel Extreme Learning Machine as kernel-based approaches and Gene Expression Programming method was assessed in breached dam peak outflow prediction. Two types of modeling were considered. First, only dam reservoir height and volume at the failure time were used as the input combinations (state 1). Then, soil characteristics were added to input combinations to investigate particularly the impact of soil characteristics (state 2). Results showed that the use of only soil characteristics did not lead to a desired accuracy; however, adding soil characteristics to input combinations (state 2) improved the models’ accuracy up to 40%. The outcome of the applied models was also compared with existing empirical equations and it was found the applied models yielded better results. Sensitivity analysis results showed that dam height had the most important role in the peak outflow prediction, while the strength parameters did not have significant impacts. Furthermore, for assessing the best-applied model dependability, uncertainty analysis was used and the results indicated that the SVM model had an allowable degree of uncertainty in peak outflow modelling.

Key words | empirical equation, gene expression programming, kernel extreme learning machine, meta model, peak outflow, uncertainty

HIGHLIGHTS

Some novelties of the present research can be summarized as follows:

- Applicability and accuracy of two different Meta models, namely, Gene Expression Programming (GEP) and Support Vector Machine (SVM) are used to predict the peak outflow from breached embankment dams.
- Two different scenarios are developed based on dam reservoir height and volume at the time of failure and soil characteristics.
- Additional to the geometrical characteristics of dam, the impact of soil characteristics on modelling the peak outflow is investigated.
- The outcomes of the SVM and GEP models are also compared with the existing empirical equations and it is shown that the intelligence models yield better results.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (http://creativecommons.org/licenses/by-nc-nd/4.0/).

INTRODUCTION

The main purpose of dam building is to enhance the quality of human life by providing drinking water, power generation, navigation, flood control and so on. However, dam failure can cause catastrophic flooding and consequently presents high risk to human life and property located in the downstream. In order to prevent and mitigate such a natural hazard, dam owners and agencies responsible for dam safety carefully study, analyze and inspect dams to identify significant failure modes. Over-topping and piping are the most encountered modes of failures causing breach of embankment dams (Wahl 2010). The breach parameters: time of failure and breach width and the peak outflow are crucial in evaluating dam risk assessments. Accurate predictions of such parameters remain a challenging task. Peak outflow rate due to dam failure is an important parameter in prediction of inundation water level for the emergency program planning and flood mitigation. The historical dam failure cases reveal a considerable range of financial, structural, and human losses. For instance, the South Fork Dam in Pennsylvania, USA, failed in 1889 due to overtopping, killing 2200 people and causing significant property losses (Singh & Scarlatos 1988). According to Wahl (2010), the accuracy of flood outflow assessment and related damages strongly depends on the proper computing of the outflow hydrograph caused by a dam break. In the past decades, accurate prediction of peak outflow has been of much interest among many investigators. Several experiments have been performed on cohesive and non-cohesive embankments to study the outflow hydrograph due to a dam break (e.g. Coleman et al. 2002; Gaucher et al. 2010). Kirkpatrick (1997) applied the data of 13 failed dams and six hypothetical failures to offer a relationship between the peak outflow and water depth. A review of the contemporary regression relationships was developed to model the breached earth dam’s peak discharge. Wahl (1998, 2004) presented a collection of the previous case studies to propose several empirical relations for estimating dam-breach incidents and the peak discharge. Pierce et al. (2010) developed the Wahl (1998) breach criteria using information about new case studies and performed regression analyses on the composite database. Duricic et al. (2013) proposed a model using the kriging approach to predict peak outflow. Hooshyaripor et al. (2015) derived statistical expressions to predict peak outflow based on observed data and generated synthetic data using a copula method. Considering the complexity and uncertainty of the dam break phenomenon, the results of the classical models are not always reliable. In fact, limited databases, untested model assumption and a lack of filed data make the predictive accuracy of these models often questionable. Finding reliable approaches for modeling the dam failure is critical for assessment of flood risk. Improvement of the quality of this modeling has been addressed by multiple research works (e.g. Thornton et al. 2011; Froehlich 2016).

A number of researchers have developed several prediction models using traditional statistical regression methods and artificial intelligence techniques. Application of various multivariate regression analyses has been a common approach in recent years for prediction of peak discharge, such as Froehlich (1995a), Pierce et al. (2010) and Thornton et al. (2011). Differential evolutionary algorithms have also been used, which can effectively solve complex optimization problems (Deng et al. 2020a, 2020b; Song et al. 2021a, 2021b).

So far, the Meta model methods [e.g. Artificial Neural Networks (ANNs), Neuro-Fuzzy models (NF), Genetic Programming (GP), Support Vector Machine (SVM), and Kernel Extreme Learning Machine (KELM)], have been used for modelling complex hydraulic and hydrologic phenomena. Some examples of the Meta model approaches’ applications are total bedload prediction (Chang et al. 2012), suspended sediment concentration prediction (Kisi et al. 2012), urban flash flood forecasting (Yan et al. 2018), and stage-discharge curves developing (Azamathulla et al. 2011). Among other Meta model approaches, SVM and GEP methods have been used in long-term prediction of lake water levels (Lan 2014), statistical downscaling of watershed precipitation (Hashmi et al. 2011), sediment transport modelling in sewer (Roushangar & Ghasempour 2017), and annual rainfall-runoff forecasting (Wang et al. 2018).

In artificial intelligence models we are looking for a learning machine capable of finding an accurate approximation of a natural phenomenon, as well as expressing it in
the form of an interpretable equation. However, this bias towards interpretability creates several new issues. The computer-generated hypotheses should take advantage of the already existing body of knowledge about the domain in question. However, the method by which we express our knowledge and make it available to a learning machine remains rather unclear (Babovic 2009). Machine learning, a branch of artificial intelligence, deals with representation and generalization using the data learning technique (Sun & Li 2020). Representation of data instances and functions evaluated on these instances are part of all machine learning systems. Generalization is the property that the system will perform well on instances of unseen data; the conditions under which this can be guaranteed are a key object of study in the subfield of computational learning theory. There is a wide variety of machine learning tasks and successful applications (Mitchell 1997). In general, the task of a machine learning algorithm can be described as follows: given a set of input variables and the associated output variable(s), the objective is learning a functional relationship for the input-output variables set. It should be noted that due to the black-box nature of artificial intelligence models, the learned relationship between the inputs and output is not revealed. These methods typically do not really represent the physics of a modeled process; they are just devices used to capture relationships between the relevant input and output variables. However, when the interrelationships among the relevant variables are poorly understood, finding the size and shape of the ultimate solution is difficult, and conventional mathematical analysis methods do not (or cannot) provide analytical solutions; these methods can predict the variable of interest with more accuracy.

Estimation of the probable flood under uncertain hydrologic conditions and routing the flood wave through downstream rivers could provide invaluable information for decision makers. However, the accuracy of assessment of flood outflow and corresponding damage are heavily reliant on the appropriate calculation of the outflow hydrograph caused by a dam break. Obviously, the higher the accuracy of the input hydrograph, the more precise outputs would result in routing models that significantly affect the ultimate risk management plan. Nowadays, Meta model techniques are widely used by many researchers for estimation of the key parameters (e.g. peak value) of the input hydrograph. Drawing upon the knowledge of the previous data mining models, this paper aims to enhance the quality of estimation of dam failure outflows using two kernel-based models (SVM and KELM) and one heuristic method (GEP) under two scenarios. In the first scenario, only dam reservoir height and volume at failure time were considered as input combinations. In the second scenario, soil characteristics were added to input combinations and the impact of soil characteristics on the peak outflow modeling was assessed. Then, the most effective variables in the prediction process were investigated using sensitivity analysis. The prediction performance of the developed models was then compared with a number of previously developed models as benchmarks. Furthermore, Monte Carlo uncertainty analysis was applied to investigate the dependability of the used models.

MATERIALS AND METHODS

Data collection and sample data set

Several data sets of well documented cases of dam failures are available in the literature. In this study, three types of data sets were used, first: 93 embankment dam failures data sets, collected from a variety of sources (i.e. Froehlich 1995a, 1995b; Wahl 1998; Xu & Zhang 2009; Pierce et al. 2010), second: the experimental data from Amini et al. (2011), and third: the data generated using Breach software. The ranges of the relevant parameters in the data sets used in this study are given in Table 1, in which $V_w$ is the volume of the water behind the dam at failure, $H_w$ is the height of the water behind the dam at failure, $Q_o$ is the peak outflow, $c$ is the cohesive strength, $\varphi$ is the internal friction angle, $d_{50}$ is the median diameter of soil material, $\gamma$ is the unit weight of the soil, and $n$ is the porosity.

Kernel-based methods

Kernel-based (KB) approaches are new methods used for classification and regression purposes. Kernel based approaches are based on statistical learning theory initiated and can be used for modeling the complex and non-linear
phenomenon. Two important kernel-based approaches are KELM and SVM which work based on the different kernel types such as linear, polynomial, radial basis function (RBF) and sigmoid functions in SVM and linear, polynomial, and RBF in KELM.

Kernel extreme learning machine (KELM)

Extreme Learning Machine (ELM) is a Single Layer Feed Forward Neural Network (SLFFNN) preparing method initially introduced by Huang et al. (2006). SLFFNN is a straight framework where information weights linked to hidden neurons and hidden layer biases are haphazardly chosen, while the weights among the hidden nodes are resolved logically. This strategy likewise has preferred execution and adapts progressively over the bygone era learning methods (Huang et al. 2006), in light of the fact that not all like traditional techniques that involve numerous variables to set up, demonstrating a complex issue utilizing this technique does not need much human intercession to accomplish ideal parameters. The standard single-layer neural system with N random information \((a_i, b_i)\), M hidden neurons, and the active function \(f(a)\) are shown as follows:

\[
\begin{align*}
\alpha_i &= [\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in}]^T \in \mathbb{R}^n \\
b_i &= [b_{i1}, b_{i2}, \ldots, b_{in}]^T \in \mathbb{R}^n \\
\sum_{i=1}^{N} \alpha_i f(w_i a_i + c_i) &= X_j, \quad j = 1, 2, \ldots, N
\end{align*}
\]

where \(w_i = [w_{i1}, w_{i2}, \ldots, w_{in}]^T\) is the weight vector that joins the input layer to the hidden layer, and \(a_i = [\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in}]^T\) is the weight vector that joins the hidden layer to the target layer. \(c_i\) shows the hidden neuron biases. The general SLFFNN network with the M hidden neuron and the activation function \(f(a)\) can predict N information with an average zero error \((\sum_{j=1}^{N} ||X_j - Y_j|| = 0)\), in which:

\[
\sum_{i=1}^{N} \alpha_i f(w_i a_i + c_i) = Y_j, \quad j = 1, 2, \ldots, N
\]

Equations (4) can be summarized as:

\[
A\alpha = K
\]

\[
A = \begin{bmatrix} f(w_1 a_1 + c_1) & \cdots & f(w_M a_1 + c_M) \\ \vdots & \ddots & \vdots \\ f(w_1 a_N + c_1) & \cdots & f(w_M a_N + c_M) \end{bmatrix}_{N \times M}
\]

\[
K = \begin{bmatrix} p_1^T \\ \vdots \\ p_N^T \end{bmatrix}_{N \times L}, \quad \alpha = \begin{bmatrix} \alpha_1^T \\ \vdots \\ \alpha_M^T \end{bmatrix}_{M \times L}
\]

The matrix \(K\) is identified as the target matrix of the hidden layers of the neural network. Huang et al. (2012) also introduced kernel functions in the design of ELM. Now number of kernel functions is used in the design of ELM such as Linear, Radial basis, Normalized polynomial, and Polynomial kernel functions. Kernel function based
ELM design is named KELM. For more detail about KELM, readers and researchers are referred to Huang et al. (2012).

Support vector machine

One of the important kernel-based approaches is SVM, which works based on the different kernel types such as linear, polynomial, radial basis function (RBF) and sigmoid functions. Such a model is capable of adapting itself to predict any variable of interest via sufficient inputs. The SVM could model non-linear decision boundaries, and there are many kernels to choose from. It is also fairly robust against overfitting, especially in high-dimensional space. A SVM constructs a hyper plane or set of hyper planes in a high or in finite dimensional space, in fact the SVM method is based on the concept of an optimal hyper plane that separates samples of two classes by considering the widest gap between two classes (see Figure 1). The appropriate selection of kernel type is the most important step in the SVM due to its direct impact on the training and classification precision. The purpose of SVM is to characterize the flattest function that has the minimum deviation from the actually obtained objectives for all training data. SVM formulation for regression aim is:

\[ f(x) = w \phi(x) + b \]  

(8)

where \( \phi(x) \) is a nonlinear function of input \( x \), \( b \) is called the bias and the vector, \( w \), is known as the weight factor. The coefficients of Equation (8) are predicted by minimizing regularized risk function as expressed below:

\[ R_{\text{min}} = C \frac{1}{N} \sum_{i=1}^{m} L_{\varepsilon}(t_i, y_i) + \frac{1}{2} ||w||^2 \]  

(9)

where

\[ L_{\varepsilon}(t_i, y_i) = \begin{cases} 0 & |t_i, y_i| \leq \varepsilon \\ |t_i, y_i| - \varepsilon & \text{Otherwise} \end{cases} \]  

(10)

\( C \): the cost factor which shows the trade-off between the weight factor and approximation error, \( \varepsilon \): the radius of the tube within which the regression function must lie, \( L_{\varepsilon}(t_i, y_i) \): the loss function in which \( y_i \) is the forecasted value, and \( t_i \) is the desired value in period \( i \). According to Equation (10), if the predicted value is out of \( \varepsilon \) tube then the loss will be the absolute value, which is the difference between the predicted value and \( \varepsilon \).

Gene expression programming (GEP)

Gene expression programming (GEP) was developed by Ferreria (2001) using fundamental principles of the Genetic Algorithms (GA) and Genetic Programming (GP). One of the benefits of GEP is the simple genetic diversity creation because genetic operators work at the chromosome level. Another benefit of GEP is its unique, multi-genetic nature, which permits the development of more complex programs consisting of several subprograms. GEP, as GA, mimics the biological evolution to make a computer program to simulate a particular phenomenon (Ferreria 2001). A GEP algorithm starts by choosing five elements including the function set, terminal set, fitness function, control parameters, and stopping condition.

The predicted values are compared with the actual values in subsequent step, and when desirable results, in accord with a priori error tolerance are found, the GEP process is terminated. If the desired accuracy is not achieved, some chromosomes are selected by roulette wheel sampling method and for obtaining new chromosomes they are mutated. After the desired fitness score is obtained, this process is terminated and then the
chromosomes are decoded for the best solution of the problem (Teodorescu & Sherwood 2008).

Performance criteria

For investigating the performance of applied methods, three statistical parameters were used: Correlation Coefficient (R), Determination Coefficient (DC), and Root Mean Square Errors (RMSE), expressions for which are as follows:

\[
DC = 1 - \frac{\sum_{i=1}^{N} (l_o - l_p)^2}{\sum_{i=1}^{N} (l_o - \bar{l}_p)^2},
\]
\[
R = \frac{\sum_{i=1}^{N} (l_o - \bar{l}_o) \times (l_p - \bar{l}_p)}{\sqrt{\sum_{i=1}^{N} (l_o - \bar{l}_o)^2 \times (l_p - \bar{l}_p)^2}},
\]
\[
RMSE = \sqrt{\frac{\sum_{i=1}^{N} (l_o - l_p)^2}{N}}
\]

(11)

where \(l_o, l_p, \bar{l}_o, \bar{l}_p, N\) are the observed values, predicted values, mean observed values, mean predicted values and number of data samples, respectively. It should be noted that predicting the parameter of interest via raw (non-normalized) data can result in undesirable predictions. So, in the current study, all input variables were scaled to fall in the range 0.1–1 to eliminate the influence of the variability of the absolute magnitudes of the data. To this end, the following equation was used.

\[
n_n = 0.1 + 0.9 \times \left( \frac{n - n_{\text{min}}}{n_{\text{max}} - n_{\text{min}}} \right)
\]

(12)

where \(n_n, n, n_{\text{max}}, n_{\text{min}}\) are the normalized, the original, the maximum and minimum amounts of parameter \(n\), respectively.

Simulation and models development

The proper choice of the input combinations is the crucial step in modelling using an intelligent approach. According to Nourani et al. (2012), dam reservoir height and volume at failure time can be considered as the important factors in dam breach peak discharge modelling. Based on Wahl (2004) and Nourani et al. (2012) the important variables which affect the peak outflow are:

\[
Q_p = f (V_w, H_w, c, \varphi, d_{50})
\]

(13)

In this study, according to Figure 2, two scenarios were considered for predicting the peak outflow. In scenario 1, the dam reservoir height and volume at failure time were used as input parameters. Models developed according to this scenario were tested using the available dam failure data in the literature and the experimental data from Amini et al. (2011). For scenario 2, soil characteristics were added to the input parameters of scenario 1 to examine the impact of soil characteristics on the models’ accuracy. Scenario 2 models were tested using 82 data generated by Breach software. Table 2 shows the models developed in this study. The used data sets were divided into two parts of training (75% of the data) and testing (25% of the data) sets (Deng et al. 2020c).

SVM, KELM, and GEP models development

Each artificial intelligence method has its own settings and parameters and for achieving the desired results, the optimized amount of these parameters should be determined. In SVM and KELM, designing the selection of the appropriate type of kernel function is important. In this section of the paper, SVM and KELM methods were evaluated using model H(III) of scenario 1 in order to select the best kernel functions of each model. The results of performance criteria considering different kernel types are shown in Figure 3. The results show that modeling with RBF kernel function results in more desirable outcomes compared to other kernel functions. Accordingly, the RBF kernel was used as the core tool for the subsequent SVM and KELM models. In using GEP, the basic arithmetic operators of (+, −, ×, /) and several mathematical functions (\(X^2, X^3, \sqrt{}, \sqrt[n]{}\)) were applied as the GEP function set. GEP models were evolved until the fitness function remained unchanged for 10,000 runs for each pre-defined number of gene. Then, the GEP model’s parameters were optimized. In Table 3 all used genetic operators are listed.
RESULTS AND DISCUSSION

Modelling based on the dam reservoir height and volume at failure time (scenario 1)

The models in the scenario 1 were developed based on available data sets of dam failures considering the $H_w$, $V_w$, or both as input variables. The defined models were tested with SVM, KELM, and GEP models to carry out the dam failure peak outflow prediction. In the modeling process, two data series were used including the available dam failures data in the literature and experimental data from Amini et al. (2011). The obtained results are shown in Figure 4(a) and listed in Table 4. Based on the statistical parameters (RMSE, R, and DC) results, it could be deduced that for both data series, between developed models the model $H(III)$, which uses both $H_w$, $V_w$ as input parameters, resulted in more desirable outcomes compared to other models in modeling of the peak outflow. This issue confirmed the importance of both dam reservoir height and volume at failure time in the peak outflow estimating process. Also, it could be seen that the impact of variable $H_w$ on the modelling process was more than variable $V_w$. The salient feature of GEP is that it can provide an explicit equation for the studied parameter. The GEP chromosomes are usually composed of one or more genes. Each gene code for a sub-expression tree (ET) and the sub-ETs interact with one another to form a more complex multi-subunit ET.

![Figure 2](http://iwaponline.com/ws/article-pdf/doi/10.2166/ws.2021.100/876463/ws2021100.pdf)

Table 2 | SVM, KELM, and GEP developed models

<table>
<thead>
<tr>
<th>Scenario 2</th>
<th>Scenario 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input variable(s)</td>
<td>Model</td>
</tr>
<tr>
<td>$H_w$, $V_w$</td>
<td>S(I)</td>
</tr>
<tr>
<td>$H_w$, $V_w$, $d_{50}$</td>
<td>S(II)</td>
</tr>
<tr>
<td>$H_w$, $V_w$, $c$, $\phi$</td>
<td>S(III)</td>
</tr>
<tr>
<td>$c$, $\phi$, $d_{50}$</td>
<td>S(IV)</td>
</tr>
<tr>
<td>$H_w$, $V_w$, $\phi$, $d_{50}$</td>
<td>S(VI)</td>
</tr>
</tbody>
</table>
Figure 4(b) shows the ETs for the model H(III). In Figure 4(b) the parameters \(d_0\) and \(d_1\) are \(V_w\) and \(H_w\), respectively. In this study, the sub-ETs were linked by addition operator. The mathematical expression of the best GEP model is as follows:

\[
Q_p = V_w + \left( \frac{H_w}{C_0} \right)^3 - \left( \frac{H_w}{C_0} \right)^4 \]  

(14)

**Modelling based on breach software data (scenario 2)**

In scenario 2, several models were defined based on dam reservoir height and volume at failure time, and soil characteristics to investigate the impact of soil properties on the performance of the models. The 82 data series extracted from breach software were used to carry out the peak outflow prediction. Table 5 and Figure 5(a) show the results of scenario 2. From the obtained results, it could be seen that between all developed models, model S(VI) with input parameters \(H_w, V_w, d_{50}, \) and \(\varphi\) performed more successfully. Also, it could be seen that model S(V) with parameters \(H_w, V_w, d_{50}, \) and \(\varphi\) approximately showed the same accuracy. According to the results of these two models, it could be indicated that cohesive strength does not have a significant impact on enhancing the models' accuracy. The obtained results showed that using only soil characteristics could not lead to the desired accuracy (model S(IV)). However, adding soil characteristics \(c, d_{50}, \) and \(\varphi\) to input combinations \((H_w, V_w)\) improve the models' accuracy. It could be seen that \(d_{50}\) increased modeling efficiency up to 5%, while \(c\) parameter increased modeling efficiency only 2%. Adding the combination of \(c\) and \(\varphi\) improved the models' accuracy up to 40%. Therefore, among soil characteristics, \(\varphi\) was more efficient. In general, it could be stated that dam reservoir height and volume at failure time and soil characteristics had impact on the peak outflow prediction.

For evaluating the impact of different parameters on the peak outflow, a sensitivity analysis was done for the SVM-best model. In this regard, the impact of each parameter was assessed by omitting it from the input set and calculating the RMSE error criteria. Figure 5(b) shows the...
sensitivity analysis results for the model S(VI). This figure clearly shows that in the peak outflow predicting $H_w$ was the most efficient parameter and the less significant parameter was $c$. Therefore, the model S(V) was selected as the superior model. The mathematical expression of GEP for model S(V) is as follows:

$$Q_p = \varphi \times (\varphi H_w + H_w - 0.179) - 0.179) + (-0.179H_w)^3 \left[\frac{H_w}{d_{50}} \times 2\varphi\right] + V_w - 0.179^3$$

(15)

Uncertainty analysis

In order to find the uncertainty of the superior SVM models, uncertainty analysis is done. According to Abbaspour et al. (2007), uncertainty is a result-dependent factor that represents the range of values a modelling result can attain.

In this study, the Monte Carlo uncertainty analysis method is used. In this procedure, two factors are used to test the robustness and analyse the model’s uncertainty. The first is the percentage of the studied outputs that are in the range of 95PPU and the second is the average distance between the upper ($X_U$) and lower ($X_L$) uncertainty bands (Noori et al. 2015). In this regard, the considered model should be developed many times (1,000 in the present study), and the empirical cumulative distribution probability of the models should be calculated. After that, the $X_L$ and $X_U$ are considered 2.5% and 97.5% probabilities of the cumulative distribution, respectively. In the proper confidence level, two important indices should be considered. First: the 95PPU band brackets most of the observations, and second: the average distance between the upper (at 97.5% level) and the lower (at 2.5% level) parts of the 95PPU be smaller than the standard deviation of the measured data (Abbaspour et al. 2007). The two mentioned
### Table 4 | Statistical parameters of developed models for scenario 1

<table>
<thead>
<tr>
<th>Model</th>
<th>Performance criteria</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R</td>
<td>DC</td>
</tr>
<tr>
<td>Dam failures data</td>
<td>SVM</td>
<td>0.732</td>
<td>0.660</td>
</tr>
<tr>
<td></td>
<td>KELM</td>
<td>0.721</td>
<td>0.648</td>
</tr>
<tr>
<td></td>
<td>GEP</td>
<td>0.706</td>
<td>0.614</td>
</tr>
<tr>
<td>H(II) SVM</td>
<td></td>
<td>0.755</td>
<td>0.702</td>
</tr>
<tr>
<td></td>
<td>KELM</td>
<td>0.744</td>
<td>0.689</td>
</tr>
<tr>
<td></td>
<td>GEP</td>
<td>0.729</td>
<td>0.667</td>
</tr>
<tr>
<td>H(III) SVM</td>
<td></td>
<td>0.891</td>
<td>0.850</td>
</tr>
<tr>
<td></td>
<td>KELM</td>
<td>0.884</td>
<td>0.841</td>
</tr>
<tr>
<td></td>
<td>GEP</td>
<td>0.883</td>
<td>0.833</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Performance criteria</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R</td>
<td>DC</td>
</tr>
<tr>
<td>Experimental data</td>
<td>SVM</td>
<td>0.659</td>
<td>0.561</td>
</tr>
<tr>
<td></td>
<td>KELM</td>
<td>0.647</td>
<td>0.554</td>
</tr>
<tr>
<td></td>
<td>GEP</td>
<td>0.635</td>
<td>0.522</td>
</tr>
<tr>
<td>H(II) SVM</td>
<td></td>
<td>0.680</td>
<td>0.597</td>
</tr>
<tr>
<td></td>
<td>KELM</td>
<td>0.671</td>
<td>0.586</td>
</tr>
<tr>
<td></td>
<td>GEP</td>
<td>0.656</td>
<td>0.567</td>
</tr>
<tr>
<td>H(III) SVM</td>
<td></td>
<td>0.752</td>
<td>0.627</td>
</tr>
<tr>
<td></td>
<td>KELM</td>
<td>0.751</td>
<td>0.627</td>
</tr>
<tr>
<td></td>
<td>GEP</td>
<td>0.755</td>
<td>0.625</td>
</tr>
</tbody>
</table>

### Table 5 | Statistical parameters of developed models for scenario 2

<table>
<thead>
<tr>
<th>Model</th>
<th>Performance criteria</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R</td>
<td>DC</td>
</tr>
<tr>
<td>S(I) SVM</td>
<td></td>
<td>0.955</td>
<td>0.893</td>
</tr>
<tr>
<td></td>
<td>KELM</td>
<td>0.941</td>
<td>0.877</td>
</tr>
<tr>
<td></td>
<td>GEP</td>
<td>0.922</td>
<td>0.848</td>
</tr>
<tr>
<td>S(II) SVM</td>
<td></td>
<td>0.961</td>
<td>0.921</td>
</tr>
<tr>
<td></td>
<td>KELM</td>
<td>0.947</td>
<td>0.904</td>
</tr>
<tr>
<td></td>
<td>GEP</td>
<td>0.927</td>
<td>0.856</td>
</tr>
<tr>
<td>S(III) SVM</td>
<td></td>
<td>0.988</td>
<td>0.967</td>
</tr>
<tr>
<td></td>
<td>KELM</td>
<td>0.973</td>
<td>0.950</td>
</tr>
<tr>
<td></td>
<td>GEP</td>
<td>0.953</td>
<td>0.899</td>
</tr>
<tr>
<td>S(IV) SVM</td>
<td></td>
<td>0.643</td>
<td>0.397</td>
</tr>
<tr>
<td></td>
<td>KELM</td>
<td>0.633</td>
<td>0.390</td>
</tr>
<tr>
<td></td>
<td>GEP</td>
<td>0.620</td>
<td>0.369</td>
</tr>
<tr>
<td>S(V) SVM</td>
<td></td>
<td>0.989</td>
<td>0.984</td>
</tr>
<tr>
<td></td>
<td>KELM</td>
<td>0.974</td>
<td>0.966</td>
</tr>
<tr>
<td></td>
<td>GEP</td>
<td>0.981</td>
<td>0.980</td>
</tr>
<tr>
<td>S(VI) SVM</td>
<td></td>
<td>0.991</td>
<td>0.983</td>
</tr>
<tr>
<td></td>
<td>KELM</td>
<td>0.976</td>
<td>0.967</td>
</tr>
<tr>
<td></td>
<td>GEP</td>
<td>0.988</td>
<td>0.981</td>
</tr>
</tbody>
</table>
Figure 5 | (a): comparison of observed and predicted dam failure peak outflow for model S(VI), (b): statistical RMSE parameter obtained from sensitivity analysis, and (c): uncertainty analysis for the SVM-best models.
indices are applied for accounting input uncertainties. For evaluating the average width of the confidence interval band, the band width indicator was suggested by Abbaspour et al. (2007) as follows:

\[
d - \text{factor} = \frac{dx}{\sigma x}
\]  
(16)

where \( \sigma x \) and \( dx \) are the standard deviation of observed data and the confidence band’s average width, respectively. The percentage of the data within the confidence band of 95% is calculated as:

\[
\text{Bracketed by 95PPU} = \frac{1}{R} \text{Cont}(j)X_L < X_{\text{reg}} < X_U
\]  
(17)

where 95PPU shows 95% predicted uncertainty; \( k \) shows the number of observed data and \( X_{\text{reg}} \) shows the current registered data. The obtained results are shown in Figure 5(c).

Based on the values obtained for the d-factor and 95PPU, it could be indicated that all observed and predicted values were within the 95PPU band for both train and test sets. Also, it was found that the amount of d-Factors for train and test datasets were smaller than the standard deviation of the observed data. Therefore, based on the results, it could be induced that the dam failure peak outflow modeling via SVM model led to an allowable degree of uncertainty.

**Comparison of the artificial intelligence models with the classical equations**

In this study, two scenarios were considered for predicting the dam failure peak outflow. In the first scenario, only dam reservoir height and volume at the time of failure were used as the inputs and results showed that the combinations of \( H_w \) and \( V_w \) led to better outcomes. Also, in the second scenario, soil characteristics were added to the combinations of \( H_w \) and \( V_w \) and results showed that the model with parameters of \( H_w \), \( V_w \), \( d_{50} \), \( c \), and \( \phi \) was the superior model. Accuracy of the best proposed models of the considered scenarios and some dam failure peak outflow semi-theoretical formulae available in literature was compared to evaluate the performance of applied approaches. The results of comparison for each scenario and testing procedure are represented in Table 6 and Figure 6. According to

**Table 6 | Comparison of prediction from proposed equations and developed models**

<table>
<thead>
<tr>
<th>Model</th>
<th>Performance criteria</th>
<th>R</th>
<th>DC</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scenario 1 data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evans</td>
<td>( Q_p = 0.72 V_w^{0.53} )</td>
<td>0.798</td>
<td>0.387</td>
<td>0.166</td>
</tr>
<tr>
<td>MacDonald and Langridge-Monopolis</td>
<td>( Q_p = 1.154(H_w V_w)^{0.412} )</td>
<td>0.779</td>
<td>0.574</td>
<td>0.138</td>
</tr>
<tr>
<td>Costa</td>
<td>( Q_p = 0.763(H_w V_w)^{0.42} )</td>
<td>0.731</td>
<td>0.419</td>
<td>0.162</td>
</tr>
<tr>
<td>Froehlich</td>
<td>( Q_p = 0.607(H_w^{24} V_w^{2.95}) )</td>
<td>0.769</td>
<td>0.218</td>
<td>0.205</td>
</tr>
<tr>
<td>SVM</td>
<td>H(III)</td>
<td>0.870</td>
<td>0.787</td>
<td>0.049</td>
</tr>
<tr>
<td>KELM</td>
<td>H(III)</td>
<td>0.871</td>
<td>0.781</td>
<td>0.051</td>
</tr>
<tr>
<td>GEP</td>
<td>H(III)</td>
<td>0.833</td>
<td>0.756</td>
<td>0.052</td>
</tr>
<tr>
<td><strong>Scenario 2 data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evans</td>
<td>( Q_p = 0.72 V_w^{0.53} )</td>
<td>0.678</td>
<td>0.352</td>
<td>0.199</td>
</tr>
<tr>
<td>MacDonald and Langridge-Monopolis</td>
<td>( Q_p = 1.154(H_w V_w)^{0.412} )</td>
<td>0.703</td>
<td>0.522</td>
<td>0.165</td>
</tr>
<tr>
<td>Costa</td>
<td>( Q_p = 0.763(H_w V_w)^{0.42} )</td>
<td>0.760</td>
<td>0.403</td>
<td>0.186</td>
</tr>
<tr>
<td>Froehlich</td>
<td>( Q_p = 0.607(H_w^{24} V_w^{2.95}) )</td>
<td>0.823</td>
<td>0.211</td>
<td>0.235</td>
</tr>
<tr>
<td>SVM</td>
<td>S(I)</td>
<td>0.925</td>
<td>0.881</td>
<td>0.105</td>
</tr>
<tr>
<td>KELM</td>
<td>S(I)</td>
<td>0.911</td>
<td>0.865</td>
<td>0.109</td>
</tr>
<tr>
<td>GEP</td>
<td>S(I)</td>
<td>0.893</td>
<td>0.838</td>
<td>0.112</td>
</tr>
</tbody>
</table>
three evaluated criteria (R, DC, RMSE), which are shown in Table 6, it could be seen that the estimated values via applied models had more accurate results than equations. It should be noted that the existing equations are developed based on special conditions, therefore the application of equations are limited to special cases of their development and did not show uniform results under different conditions. The mentioned issue can be seen in Figure 6, which showed that the obtained results from equations differ from each other and from the measured data, and had less preciseness than proposed models in this study. From the obtained results, among all equations, the MacDonald & Langridge-Monopolis equation yielded better results. It could be seen that for both scenarios, classical equations in most cases under estimated the peak outflow data. However, the obtained results by artificial intelligence models were more accurate. The intelligence applied approaches had the highest correlation with measured data and the lowest error. This issue shows the high capability of the SVM, KELM, and GEP methods in predicting the peak outflow. According to the black-box nature of artificial intelligence models, the learned relationship between input and output parameters is not apparent; however, these methods can significantly reduce the complexity of the system, identify the effective parameters and achieve the desired result.

CONCLUSIONS

Dam failure is a catastrophic phenomenon that can create many serious problems. In this paper, three data-driven predictive models, named SVM, KELM, and GEP, were used to predict the peak outflow of a breached embankment dam.
To find an effective model, two scenarios are considered in the modelling process. In scenario 1, it is considered that $Q_p$ depended on dam reservoir height and volume at the failure time. In scenario 2, soil characteristics are added to input combinations and the impact of soil characteristics was investigated on modelling the peak outflow. The results were then compared to previous research works.

From the results, it was found that scenario 2, which considered both reservoir parameters (i.e. height and volume) at the failure time and soil characteristics as the input for modelling peak outflow, led to more accurate predictions than scenario 1, which used only a combinations of reservoir height and volume parameters as inputs. And this issue confirmed the impact of soil properties on modelling peak outflow. In scenario 1, the model with input parameters $H_w$ and $V_w$ yielded better results, and in scenario 2, the model with parameters $H_w$, $V_w$, $d_{50}$, $c$, and $\varphi$ performed better than the others. However, it was found that the model with parameters $H_w$, $V_w$, $d_{50}$, and $\varphi$ approximately showed the same results and the cohesive strength parameter had a little impact on improving the efficiency of the models. Based on the obtained results, using only soil characteristics could not lead to the desired accuracy; however, adding soil characteristics ($c$, $d_{50}$, and $\varphi$) to input combinations ($H_w$ and $V_w$) improved the accuracy of models.

From the sensitivity analysis results, it was observed that $H_w$ was the most important variable in the peak outflow modelling. Also, the SVM-based model dependability was investigated using uncertainty analysis. According to the results, the SVM model had an allowable degree of uncertainty in modelling the dam failure peak outflow. A comparison was also made between artificial intelligence methods and several selected empirical formulas. The obtained results showed the superior performance of intelligence applied methods over all the semi-empirical equations in estimating the peak outflow of breached embankment dams.

**REFERENCES**


First received 20 May 2020; accepted in revised form 25 March 2021. Available online 5 April 2021.